5-2015

Big Data Efficiency, Information Waste and Lean Big Data Management: Lessons from the Smart Grid Implementation

Jacqueline Corbett
Université Laval, jacqueline.corbett@fsa.ulaval.ca

Chialin Chen
Queen's University, cchen@business.queensu.ca

R60. Big Data Efficiency, Information Waste and Lean Big Data Management: Lessons from the Smart Grid Implementation

Jacqueline Corbett
Université Laval
jacqueline.corbett@fsa.ulaval.ca

Chialin Chen
Queen's University
cchen@business.queensu.ca

Abstract
Big data has become a popular buzzword today with the underlying assumption that bigger data is better. However, by its nature, big data comes with many challenges and environmental costs. In contrast to other research that has examined the benefits and costs of big data independently, our research-in-progress provides an integrated perspective. Theoretically, we draw on the perspective of lean information management to develop the novel concept of Lean Big Data Management and a framework for evaluating it. In developing our framework, we characterize and define new concepts including Big Data Efficiency, and propose the data envelopment analysis approach as a method for measuring it. We provide an illustrative example using data from the U.S. electricity sector and discuss potential implications of this research.

Keywords
Big Data, Lean Information Management, Data Envelopment Analysis

1. Introduction
In 2011, the digital universe reached 1.8 trillion gigabytes and it is doubling every two years (Hudson 2014). Big data has become a business buzzword with the underlying assumption that bigger data is better. However, big data comes with many challenges (Jacobs 2009): organizations must adjust analytic processes, develop new data scientist capabilities, move analytics into core business functions, and ensure alignment between the business goals and technical capabilities (Davenport et al. 2012; Kiron 2013). Environmentally, big data also has substantive costs associated with powering the internet, servers, data centers and other IT infrastructures necessary for collection, analytics, storage and distribution (Hudson 2014).

Work is underway to improve the efficiency and environmental performance of data centers (Daim et al. 2009), however, the majority of research relates to IT infrastructures rather than the applications being run or data being processed. Similarly, we find gaps in the big data research. Research has examined the benefits of big data (e.g., Brynjolfsson and McAfee 2012) or the costs of big data (e.g., Winter Corporation 2013), but few studies have investigated both. Our

research-in-progress addresses these gaps by focusing on the integrated consideration and measurement of big data benefits and costs.

To inform our research, we refer to the literature on lean information management. Lean thinking is a general management philosophy focused on eliminating waste in order to optimize customer value (Liker 2003). Originally conceptualized within manufacturing, subsequent work has extended the concept to lean management (Womack and Jones 1996) and lean information management (Hicks 2007). Extending the idea further, our work develops the novel concept of Lean Big Data Management and seeks to create a conceptual framework for evaluating it. We define new concepts including Big Data Efficiency, Information Waste and Costs of Big Data. Then, we propose to illustrate these concepts using data envelopment analysis (DEA) and data from the smart grid implementation in the U.S, which is expected to result in a nine-fold increase in the amount of data available to utilities (Ambrosio 2011).

The expected contributions of our research are twofold. First, we extend the big data literature by defining the new concept of Lean Big Data Management. Our conceptualization provides a more holistic view of big data processes owing to the addition of the lean perspective which seeks to optimize complete processes rather than sub-processes where waste may exist in the gaps. Second, we propose a novel performance measure termed Big Data Efficiency to jointly measure the benefits and costs of big data management and illustrate this measure with numerical examples from smart grid implementation.

2. Literature Review: Lean Information Management

Lean information management (LIM) focuses on the creation of value and the elimination of waste throughout the information processing lifecycle (Cottyn et al. 2008). Waste includes "the additional actions and any inactivity that arise as a consequence of not providing the information consumer immediate access to an adequate amount of appropriate, accurate and up-to-date information" (Hicks 2007, p. 238). For some time, lean thinking and information technology were considered to be incompatible (Cottyn et al. 2008) because IT was seen as a barrier to lean information management (Hölttä et al. 2010). Although analogies to the seven essential sources of waste in manufacturing have been identified for information management (Table 1), their impacts have not been thoroughly investigated as they are often considered to be trivial (Hicks 2007). In contrast, we contend these costs are becoming significant in the era of big data.

3. Conceptual Development

Our conceptual framework involves four main constructs: Lean Big Data Management, Big Data Efficiency, and Costs of Big Data, and Big Data Benefits. Each of these is defined below.

We suggest the goal of *Lean Big Data Management* is the creation of a value stream related to the collection, processing, storage, analysis and mining, dissemination, maintenance and disposal of big data with minimal waste. Waste arises from different organizational processes along the value chain from collection and creation of big data to its eventual disposal. Beyond the types of waste associated with LIM, we identify new types of waste specifically associated with big data (see Table 1). Given the potential for big data waste, the question becomes: how we can measure

lean big data management and make comparisons across different organizations? Our conceptual model proposes to do this through the new construct of Big Data Efficiency.

Manufacturing Waste (Womack and Jones 1996)	**Information Management Waste**	**Big Data Waste**
Overproduction	**Flow excess** (Cottyn et al. 2008; Hicks 2007)	**Flow excess; excess data collection:** activities associated with collecting and creating data beyond value-added needs
Waiting	**Flow demand** (Cottyn et al. 2008; Hicks 2007)	**Flow demand; data congestion**: queue in data processing due to insufficient processing capacity/capability
Transport	**Unnecessary transfer** (Cottyn et al. 2008)	**Data transmission excess**: unnecessary transmission of data during collection, storage and processing
Extra processing	**Failure demand** (Cottyn et al. 2008; Hicks 2007)	**Failure demand; data cleansing waste**: resources spent on cleansing and mining "dirty data" (Hernandez and Stolfo 1998)
Inventory	**Excessive information** (Cottyn et al. 2008; Hölttä et al. 2010)	**Data storage excess; data entropy**: over-time the usability of data diminishes (Christensen et al. 2011)
Motion	**Incompatibility** (Cottyn et al. 2008; Hölttä et al. 2010)	**Data integration failure**: extra processing and data integration efforts taken to accommodate inefficient layout, defects, reprocessing, and excess data
Defects	**Flawed flow** (Hicks 2007; Hölttä et al. 2010)	**Flawed flow; flawed data waste**: activities resulting from poor data quality (Olson 2003)

Table 1: Sources of waste

Based on the input-output model in economics (Miller and Blair 2009), we define *Big Data Efficiency* as the ratio of the total output to total inputs of a big data management system. The total input may include different forms of quantifiable resources, such as costs, time and energy, for collecting, distributing, compiling, sorting, mining, processing and maintaining big data. The total output is measured by different quantifiable benefits created from analyzing big data, such as improved profitability, service quality and environmental performance. Big Data Efficiency measures how efficiently resources are used for value creation.

The *costs of big data* can be categorized according to operational costs and environmental costs. Examples of operational costs associated with big data include (but are not limited to):

- *Systems costs*: Costs to acquire, maintain, support and upgrade hardware and software plus the cost of space, power, and cooling (Winter Corporation 2013).
- *System and data administration costs*: Costs of expert staff to administer the system and the data stores (Winter Corporation 2013).
- *Integration costs*: Costs of developing or acquiring ETL (extract, transform and load) functionality to prepare data for analytic use; developing processes to cleanse the source data, recognize it as necessary, and store it in accordance with an integrated database design (Winter Corporation 2013).
- *Querying and analytical costs*: Costs associated with developing queries that can be expressed in SQL, or the development of procedural programs that demand data analyses too complex to express in SQL (Winter Corporation 2013).

- *Application costs*: Costs of developing applications using data to support repeatable processes (Winter Corporation 2013), such as demand-side management.

There are also important environmental costs, deriving primarily from the energy consumption related to the operational costs of big data. Energy and environmental costs for powering the internet, servers, data centers, and other IT devices and infrastructures to collect, distribute, compile, store, mine, process and maintain big data (Coroama and Hilty 2014) are commonly measured through energy consumption and carbon emissions.

Recently, there has been much hype around the *benefits of big data* to organizations, with some predicting up to 60% improvements in operating margins and enhanced competitive advantage (McGuire et al. 2012). Improvements to operating efficiency result from greater and faster access to information to support organizational decision-making (Brynjolfsson and McAfee 2012). Big data coming from GPS and mobile devices can also contribute to time and fuel savings (McGuire et al. 2012). In the electricity sector, the big data is expected to transform the grid into a more reliable, affordable, efficient and environmentally benign supply system (Fox-Penner 2010).

4. Methodology

For this research, we propose to use data envelopment analysis (DEA). DEA is a nonparametric method for multi-objective performance evaluation through benchmarking (Charnes et al. 1978). DEA is commonly used to measure and characterize how the limited resources (inputs) are utilized to achieve multiple objectives (outputs). As a nonparametric method, DEA has the advantage of uncovering the relationships among multiple objectives without the use of any explicit mathematical function. Another advantage of DEA is the capability to synchronize different input/output measurements on a consistent basis to avoid potential issues associated with scaling and different measurement units (Zhu and Cook 2013).

Our research setting is the smart grid. Through the implementation of smart meters and advanced metering infrastructures (AMI), electricity service providers ("utilities") are entering the era big data. We will test our model in the context of the U.S. electrical utility sector, comprising over 3200 utilities. There are substantial variations among utilities in terms of size, ownership, geography and progress in implementing the smart grid, providing a rich research context. More practically, there is a readily-available and consistent data set from the U.S. Energy Information Administration (EIA). We focus on the utilities' implementation of AMI and their demand-side management (DSM) activities. To illustrate the operationalization of our constructs and to demonstrate the DEA methodology, we provide a simple example below.

5. A Numerical Example of Big Data Efficiency

To determine Big Data Efficiency, we require measures for the inputs and outputs for each decision-making unit (utility). In this illustration, we include two measures for inputs: number of AMI devices and indirect costs associated with DSM. In the absence of knowing specific operational costs, the number of AMI devices provides a reasonable measure for system and data administration costs. Similarly, for environmental costs, the volume of electricity consumption is likely to increase with the number of AMI devices. There is no need to determine the actual environmental impact because we are concerned with the relationship (efficiency), rather than

absolute levels. The second input variable is the indirect costs of utilities' DSM programs. According to the EIA, these costs are attributable to cost categories such as administration, monitoring and evaluation. Although this measure includes indirect costs which may not relate specifically to big data, it represents a reasonable proxy for application-type costs associated big data. For the outputs, we include two performance measures: energy efficiency effects of DSM programs; and load balancing as measured by actual reductions in annual peak load achieved by consumers participating in DSM programs (Energy Information Association 2011).

As our data extraction and compilation is not completed, we used a small sample of 10 hypothetical utilities. All input and output values were randomly generated based on the average values across relevant utilities in the 2010 EIA Annual Electric Power Industry Report. The CRS (Constant Returns to Scale) Multiplier Model of DEA was used in the analysis. The inputs, outputs and the Big Data Efficiency scores are shown in Table 2.

From this illustration we can highlight certain key results. First, we consider Utility 9 which has the lowest efficiency score of 0.36169. This case may present a typical example of over-investment in big data with high operational costs (high costs for powering a large number of AMI units) which cannot be justified by the relatively low outputs (especially energy efficiency).

Alternatively, we find Utilities 3, 4, 7, 8, and 10 all of which have big data efficiency scores of 100%. A utility with 100% efficiency represents a decision making unit with the best performance along a particular direction (i.e., a particular combination of inputs and outputs) in the multi-dimensional space. The efficient units (with 100% efficiency) form the so-called 'best-practice frontier' in the multi-dimensional space. Thus, these cases are examples of efficient units which achieve the best performances in energy efficiency and/or load balance with mostly low to moderate levels of investments on inputs.

	Inputs		**Output**		
	Indirect Cost	**AMI Units**	**Energy Efficiency**	**Load Balance**	**Efficiency**
Utility 1	58 442 000	98 106	12 123	6 118	0.72768
Utility 2	58 776 858	141 756	15 878	5 981	0.75995
Utility 3	37 413 630	102 055	13 987	8 844	1.00000
Utility 4	75 578 196	58 796	16 559	3 468	1.00000
Utility 5	43 815 236	105 084	6 110	3 209	0.39314
Utility 6	75 151 826	116 862	14 635	4 423	0.67245
Utility 7	31 713 243	139 462	13 961	3 734	1.00000
Utility 8	32 090 938	65 719	10 728	4 601	1.00000
Utility 9	76 720 925	142 098	6 434	6 698	0.36169
Utility 10	29 770 352	66 613	8 248	8 724	1.00000

Table 2: Numerical Example of DEA Analysis

6. Conclusion

In this paper, we propose that Big data Efficiency can be used as an integrated measurement of the costs and benefits in Lean Big Data Management. This integrated model and measure

represent a main contribution of this work to the literature. To date, there has been limited research of lean management in the context of information systems (Bortolotti and Romano 2012), yet we believe the integration of lean with IS provides a fruitful new avenue for considering the performance of data-intensive IS solutions.

Our work also has important implications for practice. First, Big Data Efficiency can be achieved when any data or data management activities that deviate from the true analytical needs are considered information waste and should be eliminated. By reducing big data wastes, the financial and environmental performance of organizations should improve. Second, we suggest Lean Big Data Management can be realized with a pull-based system (Cottyn et al. 2008) in which data collection is driven by the true analytical needs of data consumers. Current models for big data tend to rely on push-based systems (Franks 2012). There has been initial consideration of more efficient pull-based systems for the electricity sector (Tsigkas 2011) and our work suggests continued efforts in this area may be worthwhile.

References

Ambrosio, R. (2011) "Managing the Information Glut." Retrieved May 13, 2011, from http://www.environmentalleader.com/2011/05/12/managing-the-information-glut/

Bortolotti, T. and P. Romano. (2012) "'Lean First, Then Automate': A Framework for Process Improvement in Pure Service Companies. A Case Study," *Production Planning & Control: The Management of Operations* (23)7, pp. 513-522.

Brynjolfsson, E. and A.P. McAfee. (2012) "Winning the Race with Ever-Smarter Machines," *MIT Sloan Management Review* (53)2, pp. 53-60.

Charnes, A., W.W. Cooper and E. Rhodes. (1978) "Measuring the Efficiency of Decision Making Units," *European Journal of Operational Research* (2), pp. 429-444.

Christensen, S.W., C.C. Brandt and M.K. McCraken. (2011) "Importance of Data Management in a Long-Term Biological Monitoring Program," *Environmental Management* (47), pp. 1112-1124.

Coroama, V.C. and L.M. Hilty. (2014) "Assessing Internet Energy Intensity: A Review of Methods and Results," *Environmental Impact Assessment Review* (45), pp. 63-68.

Cottyn, J., K. Stockman and H. Van Landegnhem. (2008). "The Complementarity of Lean Thinking and the ISA 95 Standard," in: *WBF 2008 European Conference*. Barcelona, Spain: pp. 1-8.

Daim, T., J. Justice, M. Krampits and M. Letts. (2009) "Data Center Metrics: An Energy Efficiency Model for Information Technology Managers," *Management of Environmental Quality: An International Journal* (20)6, pp. 712-731.

Davenport, T.H., P. Barth and R. Bean. (2012) "How 'Big Data' Is Different," *MIT Sloan Management Review* (54)1, pp. 43-46.

Energy Information Association. (2011). "Annual Electric Power Industry Report Instructions." Energy Information Association.

Fox-Penner, P. (2010) *Smart Power: Climate Change, the Smart Grid, and the Future of Electric Utilities*. Washington, DC: Island Press.

Franks, B. (2012) *Taming the Big Data Tidal Wave*. Hoboken, NJ: John Wiley & Sons, Inc.

Hernandez, M.A. and S.J. Stolfo. (1998) "Real-World Data Is Dirty: Data Cleansing and the Merge/Purge Problem," *Data Mining and Knowledge Discovery* (2)9, pp. 9-37.

Hicks, B.J. (2007) "Lean Information Management: Understanding and Eliminating Waste," *International Journal of Information Management* (27), pp. 233-249.
Hölttä, V., K. Mahlamäki, T. Eisto and M. Ström. (2010) "Lean Information Management Model for Engineering Changes," *World Academy of Science, Engineering and Technology* (4), pp. 1213-1220.
Hudson, G. (2014) "How Much Energy Does the Internet Use?" Retrieved April 9, 2014, from http://cleantechnica.com
Jacobs, A. (2009) "The Pathologies of Big Data," *Communications of the ACM* (52)8, pp. 36-44.
Kiron, D. (2013) "Organizational Alignment Is Key to Big Data Success," *MIT Sloan Management Review* (54)3, pp. 1-6.
Liker, J. (2003) *The Toyota Way*. New York: McGraw-Hill.
McGuire, T., J. Manyika and M. Chui. (2012) "Why Big Data Is the New Competitive Advantage," *Ivey Business Journal* (July/August).
Miller, R.E. and P.D. Blair. (2009) *Input-Output Analysis: Foundations and Extensions*, (2nd ed.). Cambridge, UK: Cambridge University Press.
Olson, J. (2003) *Data Quality: The Accuracy Dimension*. San Francisco, CA: Morgan Kauffman Publishers.
Tsigkas, A. (2011) "Open Lean Electricity Supply Communities: A Paradigm Shift for Mass Customizing Electricity Markets," *Enegy Systems* (2)3-4, pp. 407-422.
Winter Corporation. (2013) *Big Data: What Does It Really Cost?* Cambridge, MA: Winter Corporation.
Womack, J.P. and D.T. Jones. (1996) *Lean Thinking: Banish Waste and Create Wealth within Your Corporation*. London: Simon and Schuster.
Zhu, J. and W.D. Cook. (2013) *Data Envelopment Analysis: Balanced Benchmarking*. CreateSpace Independent Publishing.